Uncertainty in Bus Arrival Time Predictions: Treating Heteroscedasticity With a Metamodel Approach

O'Sullivan, Aidan; Pereira, Francisco Camara; Zhao, Jinhua; Koutsopoulos, Harilaos N.

Published in:
I E E E Transactions on Intelligent Transportation Systems

Link to article, DOI:
10.1109/TITS.2016.2547184

Publication date:
2016

Document Version
Peer reviewed version

Uncertainty in Bus Arrival Time Predictions: Treating Heteroscedasticity with a Meta-Model Approach

Aidan O'Sullivan, Francisco C. Pereira, Jinhua Zhao, and Harilaos Koutsopoulos

Abstract—Arrival time predictions for the next available bus or train are a key component of modern Traveller Information Systems (TIS). A great deal of research has been conducted within the ITS community developing an assortment of different algorithms that seek to increase the accuracy of these predictions. However, the inherent stochastic and non-linear nature of these systems, particularly in the case of bus transport, means that these predictions suffer from variable sources of error, stemming from variations in weather conditions, bus bunching and numerous other sources. In this paper we tackle the issue of uncertainty in bus arrival time predictions using an alternative approach. Rather than endeavour to develop a superior method for prediction we take existing predictions from a TIS and treat the algorithm generating them as a black box. The presence of heteroscedasticity in the predictions is demonstrated and then a meta-model approach deployed that augments existing predictive systems using quantile regression to place bounds on the associated error. As a case study this approach is applied to data from a real-world TIS in Boston. This method allows bounds on the predicted arrival time to be estimated, which give a measure of the uncertainty associated with the individual predictions. This represents to the best of our knowledge the first application of methods to handle the uncertainty in bus arrival times that explicitly takes into account the inherent heteroscedasticity. The meta-model approach is agnostic to the process generating the predictions which ensures the methodology is implementable in any system.

Index Terms—Intelligent Transportation Systems, bus arrival time predictions, quantile regression, heteroscedasticity, Gaussian process.

1 Introduction

Efforts to increase the use of public transport in urban areas, as part of a strategy to reduce traffic congestion and the associated problems such as pollution and poor air quality, have led public transport authorities of varying sizes to invest in advanced Traveller Information Systems (TIS). These systems aim to go beyond static schedule information and provide commuters with access to real-time information on the status of bus or rail services to allow them to better plan their journeys. Transport authorities see the benefits realised from deploying real-time bus arrival information systems as; improved customer service, increased customer satisfaction and convenience and greater visibility of transit in the community. One of the perceptions among customers is that bus services have improved and that people traveling late at night now have the reassurance that the next bus is not far away [1]. A number of studies have been conducted to evaluate these benefits and report both a significant though small increase in ridership in before and after studies [2], [3] as well as increased feelings of passenger safety when traveling at night [4].

- *A. O'Sullivan is with the Future Urban Mobility group, Singapore MIT Alliance for Research and Technology, Singapore. E-mail: aidanosull@gmail.com*
- *F. Pereira is with the Technical University of Denmark. J. Zhao is with the Massachusetts Institute of Technology and H. Koutsopoulos is with Northeastern University .*

Manuscript received August 10, 2015; revised March 16, 2016.

The bus arrival time predictions provided by these systems are an example of the travel time prediction problem, where the goal is to accurately estimate the time taken for a bus to travel from it's current location to the stop location. These arrival time predictions are made possible by the deployment of Global Positioning System (GPS) based Automatic Vehicle Location (AVL) technology, which was originally intended to increase operational efficiency through better monitoring and controlling of vehicle fleets [1]. As these deployments matured the potential for this data was recognised and a number of algorithms developed to make travel time predictions from this data. These algorithms are typically based on a kalman filter to predict the time to arrive at a location based on the current location and speed in combination with historical data, however there are a number of proprietary technologies also in use and increasingly machine learning techniques such as neural networks are being used [5]. For a review of the current state of the art see [6], [7], [8].

In this paper we study the uncertainty in bus arrival time predictions treating the algorithm making predictions as a black-box and making use of data from a real world TIS in Boston. We emphasise that the focus of this work is not on improving the accuracy of the point predictions. Rather our objective is to capture the uncertainty associated as the reliability of these predictions is an area that is often not properly analysed [9]. This is a key issue for travelers as accurate travel time predictions reduce the uncertainty in decision making about departure time and route choice which in turn reduce stress and anxiety [10]. Indeed it has been found

that the reliability of travel times are valued just as much or even more than improvements in the average travel time [10], [11], [12]. The prediction problem is complicated as bus travel times are the result of nonlinear and complex interactions of many different constituent factors influencing either demand (e.g. passenger's demand or traffic flow) or capacity (e.g. accidents, weather conditions, route characteristics) [13]. The probabilistic nature of some determining factors such as passenger demand, bus drivers' behaviour, traffic accidents and more importantly signal delay experienced by different buses leads to stochastic behaviours in the system [14]. In order to gain an understanding of the uncertainty of travel time predictions in such a system it is necessary to consider the size and dynamics of the associated variance. However as discussed in [9] these are generally disregarded entirely or taken as constant. This constant variance assumption is known as homoscedasticity and implies that the reliability of all the predictions are identical. Intuitively, and as we shall demonstrate empirically, this assumption is inappropriate, the expected value of the error terms may not be equal and the error terms may reasonably be expected to be larger for some points or ranges of the data than others. This behaviour is known as heteroscedasticity. The presence of heteroscedasticity means the reporting of a single, usually mean, travel time value misses out key information.

A more meaningful approach is to consider travel time prediction as an example of probabilistic inference which naturally leads to predicting the most probable distribution of travel time, rather than one crisp value. One approach in this vein has been to use models capable of modelling this evolving prediction variance and throw away values that suffer from a value above a threshold and deemed unreliable as proposed in [9]. Recent work analysing arrival time predictions for a bus route in Dublin advocated completely disregarding any prediction made from over 6km away [15] due to the associated lack of reliability. An alternative approach which we follow here is the use of Prediction Intervals (PIs) which take into account both the uncertainty in model structure and noise in input data. PIs have recently attracted some interest in the transportation field for travel time modelling using neural networks [5], [14]. In this work we take a different approach to constructing the PIs using quantile regression [16], which makes no assumptions as to the structure of the error process, as a post processing 'meta-model' approach that can be applied to an existing predictive system. This approach was introduced in [17] and applied to a simple, historical dataset of highway speed/density data to construct PIs. In this paper we further develop this approach using a more sophisticated model and applying it to a ubiquitous real-time service with potential direct impact to the public.

This paper makes two main contributions to the field. Firstly the presence of heteroscedasticity in arrival time predictions is demonstrated empirically using data from a real-world TIS. Secondly, having motivated our work in this manner, we develop a black box solution that can be applied to any prediction system using the error between the predicted and observed arrival times to estimate PIs associated with the predictions made. This approach is a considerable methodological advance to the earlier meta-model presented in [17], as now we use a Gaussian Process, which demonstrates to be a more powerful proposal than cubic splines.

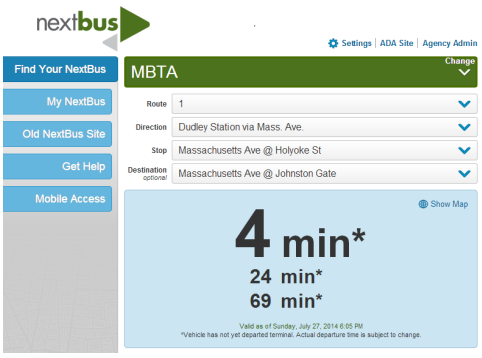


Fig. 1: Nextbus arrival predictions are provided on their website in the form of a single valued time, [18].

The paper is organised as follows, the next section introduces the data which will be used throughout the paper consisting of arrival time predictions for two bus routes in Boston. The presence of heteroscedasticity in the predictions is demonstrated rigorously and the following sections introduce our approach to address the uncertainty that arrives as a result. Section 3.1 introduces PIs in more detail and the associated evaluation metrics that will be used to assess performance. Following this we give some background on the method of quantile regression with some details on model estimation in Section 3.2. The results of applying our meta-model approach to the data are described in Section 4. The paper concludes with a discussion of some of the key aspects of this work and directions for future work.

2 Data

The data sets utilised were obtained from an arrival time prediction service, Nextbus [18], for two bus routes in the city of Boston. The main objective of this section is to describe both routes in detail and demonstrate the presence of heteroscedasticity in the errors between predicted and observed arrival times. The bus routes studied are the route 1 bus operated by the Massachusetts Bay Transport Agency (MBTA) and the MIT Boston daytime shuttle bus a privately run service for MIT staff and students. The routes were chosen for their different characteristics, one being a popular main line route while the other consists of a single vehicle on loop and therefore represents the simplest possible scenario. Despite the simplicity of the second route the predictions still were found to exhibit heteroscedasticity as will be demonstrated in this section. In both cases the predicted arrival times were obtained from Nextbus. Nextbus is a company that provides real-time passenger information systems for many major transport organisations across over 30 states in the United States and Canada. Their predictions are based on GPS and AVL data to track the vehicle in transit and estimate predicted arrival times based on it's current position. Commuters can access these predictions via a smartphone application or through the Nextbus website to see when the next bus will arrive. The information is provided in the form of a time, e.g. 5 minutes till the next arrival, a typical example for route 1 is shown in Figure 1.

While the service does offer predictions for the second and third buses to arrive we restrict our attention to predictions for the next bus to arrive in order to avoid problems of bus identification. The error between predicted and observed time of arrival is calculated as

$$e = \hat{y} - y$$

so a positive error indicates that the bus arrived before the predicted time and a negative error indicates that the bus arrived later than the predicted time.

2.1 MBTA Route 1

The first service analysed is the Route 1 bus in Boston operated by the Massachusetts Bay Transport Agency, the public transport agency for the area. The route was chosen as it is one of the main and most used routes in the area. Three months of data from March to May 2014 was obtained with an observation occurring every thirty seconds from 9am to 7pm, which equates to approximately 70000 observations. The route is shown in Figure 2a and connects Harvard square to Dudley station passing through one of the Cambridge area's busiest streets Massachusetts Avenue and traversing Harvard bridge. The typical journey time is around 40 minutes and bus frequency is approximately one every 10 mins.

2.2 MIT Shuttlebus

The second route studied is the MIT Boston daytime shuttlebus which consists of a single vehicle operating on a continuous loop that is scheduled to arrive every 25 minutes. The route is made up of eight stops covering close to five miles of the city, and these are indicated in Figure 2b. Predictions are made with respect to arrivals at stop 77 mass ave, stop 7. As before the data is split into training and test set, the training set consists of a week's worth of Nextbus predictions which are generated every 30 seconds and the actual arrival times. The test set provides the same data for a single day. This route is operated by a single vehicle which makes it the simplest possible setting, without issues of bus bunching etc., however as we will demonstrate in the next section there is still evidence of heteroscedasticity in the error in predictions

2.3 Tests for heteroscedasticity

In this section we study the data for evidence that the arrival predictions are subject to varying variances, heteroscedasticity. In order to test for the presence of heteroscedasticity, preliminary analysis was carried out using some standard statistical tests. To begin with the mean, median, skewness and kurtosis of the distributions of the residuals were calculated for both datasets, and the values for these are shown in Table 1. In both datasets different values are obtained for both the mean and median and there is a bias towards the negative direction indicating a tendency to predict an arrival time that is earlier than the true observed arrival time. The MBTA residuals are negatively skewed and leptokurtic, (positive kurtosis), meaning acutely peaked with fat tails. The shuttlebus residuals are positively skewed and highly leptokurtic with a very high positive value. The residuals are plotted in Figure 3 and as can be seen in combination with the values of Table 1 the distributions are highly non-Gaussian. However this in itself does not provide sufficient evidence of heteroscedasticity. Traditionally the Breusch-Pagan [19] test and the White's test [20], are used to test the hypothesis of non-constant variance. However both these tests rely on knowledge of the model used to generate the predictions and access to the covariates or predictor variables. Our 'black-box' approach is agnostic to the process generating the predictions and the features used so these tests are not applicable, instead we utilise the Fligner-Killeen test [21]. This is a non-parametric test that is very robust to departures from normality. The statistic tests for homogeneity between groups, therefore we must divide our predictions into groups to compare the variance in the error between those made from short term and long term. The data was binned into groups such that an equal number of samples were in each group with thresholds at 500 seconds, 1000 seconds, 1500 seconds and above. The results of the tests on both datasets were that the variance of the error between groups were significantly different with p-values less than 0.001 returned. This result was supplemented with evidence from Engle's Autoregressive Conditional Heteroscedasticity test [22]. The test procedure is to regress the squared residuals on a constant and q lags and test the null hypothesis that the correlation between lags is equal to zero, the test statistic follows a χ^2 distribution and the associated p-value can be obtained. For both datasets a p-value less than 0.001 was obtained indicating that the null hypothesis of constant variance can be rejected and that the variance of the error is non-constant.

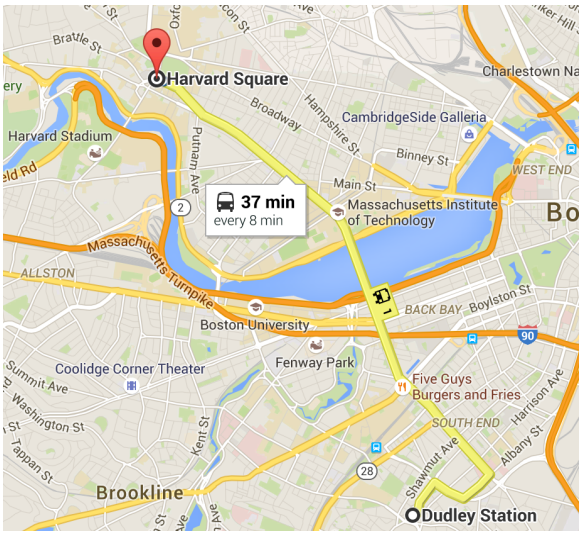
Having demonstrated the presence of heteroscedasticity in our bus arrival time prediction data appropriate methods for handling this characteristic uncertainty are required. Clearly assuming the predictions are Gaussian distributed and reporting a mean value is inappropriate and rather a prediction interval with upper and lower bounds is preferred which will be discussed in the next section.

3 Methodology

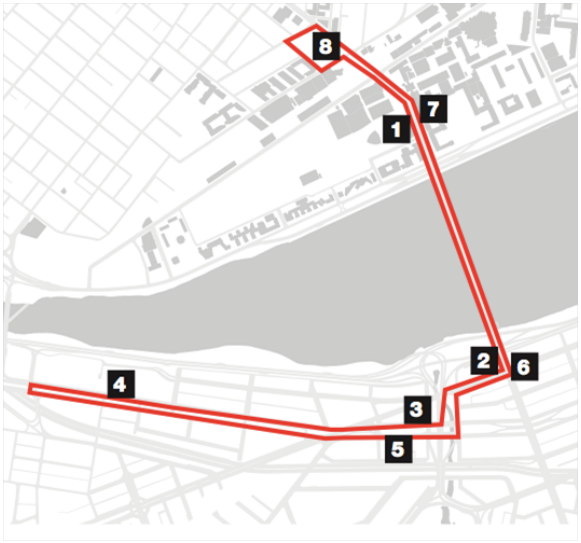
This section describes our approach to handling the demonstrated heteroscedasticity in the bus arrival time predictions. Quantile regression is used to construct upper and lower bounds on the associated prediction creating a prediction interval. As the behaviour is non-linear, the error does not increase linearly with the prediction, flexible functional forms are required to estimate the quantiles and we utilise a Gaussian Process quantile regression model. This type of model has not previously been deployed in such a setting and a splines based approach is used to compare the benefits

Measure	MBTA	ShuttleBus
Mean	-71.41	-48.71
Median	-38.00	-33.00
Skewness	-0.44	1.78
Kurtosis	4.90	15.28

TABLE 1: Moments of residuals. The measures calculated for the residuals of the MBTA and MIT ShuttleBus predictions demonstrate that the distributions are far from Gaussian.

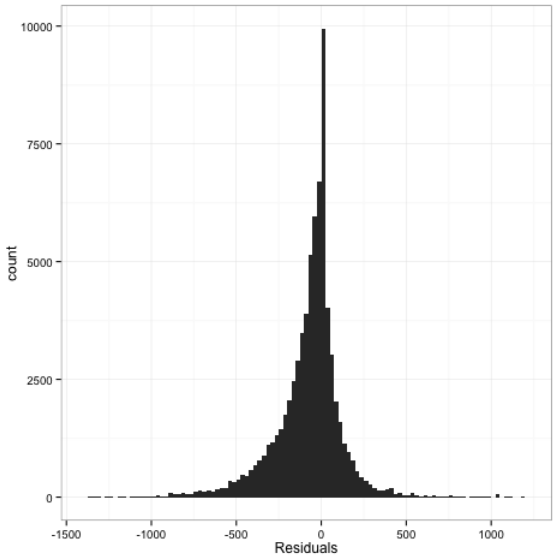


(a) MBTA bus 1 route.

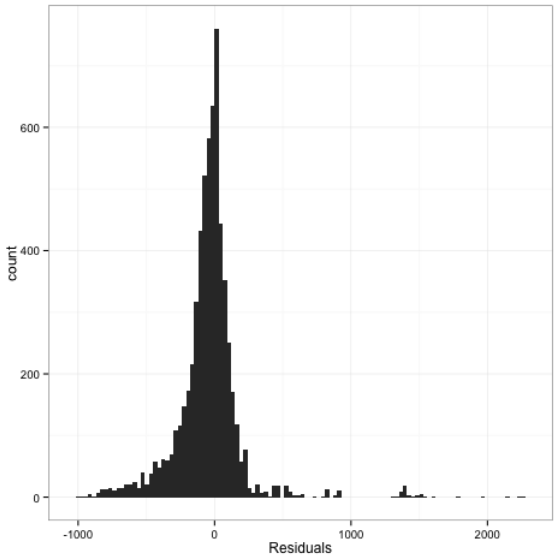


(b) MIT shuttlebus route.

Fig. 2: The bus routes our arrival time predictions are made for, Figure 2a MBTA route 1 and Figure2b the MIT shuttle bus. Route 1 takes approximately 40 minutes and connects Harvard square and Dudley station. Figure2b shows the looped route operated by the shuttle bus, predictions are made with respect to stop 7.



(a) MBTA residuals



(b) MIT shuttlebus residuals

Fig. 3: Distibution of residuals for MBTA route 1, Figure 3a, and MIT ShuttleBus, Figure 3b. The residual errors from the predictions are highly non-Gaussian.

of the more complex and computationally intensive GP model. We now provide some background on prediction intervals.

3.1 Prediction Intervals

Given the demonstrated uncertainty in arrival time predictions it can be considered more appropriate to provide travellers with upper and lower bounds on the possible arrival times rather than a single value. This is an idea that has been advanced using confidence intervals [10] and also prediction intervals [5] we will elaborate on the difference between the two which is subtle while defining PIs. A PI, with a confidence level of $(1-\alpha)\%$, is defined as a random interval developed based on past observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ for future observations:

$$\text{PI} = [L(x), U(x)],$$

such that:

$$P(L(\boldsymbol{x}) \leq x_{n+1} \leq U(\boldsymbol{x})) = 1 - \alpha,$$

where $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ correspond to the lower and upper bounds of PIs. The confidence level $(1-\alpha)\%$ of a PI refers to the expected probability that the real value is within the predicted interval. In the arrival time prediction setting

a PI can be thought of as a window within which we expect the vehicle to arrive with some chosen probability. It is important to note the distinction between a prediction interval and a confidence interval obtained from using the mean value and an estimate of the variance. The prediction interval creates a window within which we expect the next predicted value to fall with some probability, e.g. 90% of the time, whereas a confidence interval gives us a range within we would expect to find the **mean** value with some probability.

From the above definition it's clear that we can create a PI that contains the bus arrival time with 100% of the time by making the interval arbitrarily large. However informing a traveler that their bus will arrive with 100% certainty within the next 6 hours is not particularly useful. This gives us some insight as to how to assess the performance of estimated PIs. [14] defined the following metrics which evaluate both the length and coverage probability of the predicted interval, Probability Interval Coverage Probability:

$$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} c_i,$$

where $c_i = 1$ if $y_i \in [y^{\tau-}, y^{\tau+}]$ and Mean Predicted Interval Length:

$$\text{MPIL} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\tau+} - y_i^{\tau-}\right),$$

where τ^+ indicates the upper bound and τ^- the lower bound. The goal is to get as close as possible to the desired coverage probability with the smallest MPIL.

We construct our PI using quantile regression which is introduced in the next section, to estimate independent upper and lower bounds on the error between the observed arrival time and that which was predicted. Given the complex non-linear nature of such a signal we deploy a Gaussian Process (GP) [23] for model estimation.

3.2 Quantile Regression

Traditionally, regression methods, studying the relationship between a target variable, y, and a set of predictor variables, $\boldsymbol{x}$, have been dominated by the least squares approach which constructs a regression that minimises the sum of squared residuals [24]. This least squares approach has a number of convenient properties that have contributed to it's prominence, it is computationally and conceptually straight forward [25], captures the conditional mean of the target variable given the predictor variables and is optimal under the condition of constant Gaussian noise.

While the mean of the function may be a good way to summarise the relationship in general, consideration of other loss functions can allow us to extract a more complete picture of the relationships which may be more appropriate for different applications. Minimising a sum of asymmetrically weighted absolute residuals, also known as the tilted or pinball loss function, leads to a regression on the quantiles [16], [26]. This Quantile Regression (QR) has found use in many fields [26] from econometrics [27] to epidemiology [28] and more recently Big Data applications [29]. QR makes no assumptions about the nature of the error process, as opposed to the assumption of constant Gaussian error in ordinary least squares linear regression, making it a semi-parametric method.
The tilted loss function is defined as:

$$L_\tau(y - y^*) = \left\{ \begin{array}{ll} \tau(y - y^*), & y \geq y^*, \\ (1-\tau)(y - y^*), & y < y^* \end{array} \right\}, \qquad (1)$$

where $\tau \in [0, 1]$ defines the asymmetry point, for example $\tau = .5$ gives a median regression. Linear programming methods are required to solve the minimisation problem as it stands and directly obtain the desired quantiles, however this minimisation is exactly equivalent to the maximisation of a likelihood function formed by combining independently distributed Asymmetric Laplace Distributions (ALD) [30]. This has opened up QR to a Bayesian treatment which has been a rich area of research in the last decade with numerous alternate models and estimation methods developed. In this paper we take a Bayesian non-parametric approach using a GP to estimate the desired quantile fucntions as in [32]. GPs represent one of the most popular and advanced methods in the current state of the art for regression. Their popularity stems from the flexibility of the method which can be thought of as an infinite dimensional multivariate Gaussian distribution [23] which gives us a very powerful model to estimate the quantiles. We briefly outline the GPQR approach as follows but for further detail see [32]. The training of the model proceeds by maximising a utility function based on the ALD. The utility function is defined as:

$$\mathcal{U}_\tau(\boldsymbol{y}, \boldsymbol{q}) = Z\exp\left[-\sum_{i=1}^{N} L_\tau(y_i, q_i)\right],$$

where $\boldsymbol{q}$ is the predicted value of the τ quantile, $\boldsymbol{y}$ the observations, Z the normalisation constant and L_τ is the ALD for quantile τ given by:

$$L(t|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma}\exp\left[-\frac{t-\mu}{\sigma}(\tau - I(t \leq \mu)\right],$$

where $I(t \leq \mu)$ is the indicator function which is 1 if the condition is true. A GP prior is placed on the quantile regression function:

$$p(\boldsymbol{q}) = \mathcal{GP}(\boldsymbol{q}|0, K),$$

and the model is then trained by maximising the integral:

$$\arg\max \int_q \mathcal{U}_\tau(\boldsymbol{y}, \boldsymbol{q}) p(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}.$$

This integral is analytically intractable however it can be locally approximated using an Expectation Propagation algorithm outlined in [33]. The hyperparameters associated with the GP are learned in the same fashion as ordinary GP regression [23]

The numerous contributing sources of uncertainty affecting arrival time predictions, outlined in Section 1, leads to a stochastic process that is highly complex and motivates the use of a highly flexible method capable of capturing such behaviour and this motivates our use of the GPQR. However we will also compare the results obtained with GPQR to a commonly used alternative QR model to ascertain

whether the additional complexity is justified by improved performance. To demonstrate the method we apply it to the real world data introduced in Section 2 and evaluate the results in the next section.

4 EXPERIMENTS

Section 2 introduced two real world datasets of bus arrival time predictions for MBTA route 1 and the MIT shuttle bus route. In both data sets the presence of heteroscedasticity in the residuals between predictions and observed errors was rigorously demonstrated. In this section we evaluate the performance of the methods outlined in Section 3, as applied to this data to handle the inherent heteroscedasticity and compare with a homoscedastic model to illustrate the advantages of our approach. In both data sets we will divide the data into a training set used to learn the model and a test set to assess performance. The homoscedastic approach will utilise a linear splines regression to estimate the median error between predicted and observed arrival time and build a 90% confidence interval using the variance of this model. This is compared to a PI constructed with an upper and lower bound learned using splines quantile regression. These models were simulated in the R statistical programming language [34] using the package *quantreg* [35]. The more advanced Gaussian Process quantile regression approach will be compared with the splines QR to ascertain whether the increased complexity of the GP results in improved performance. This was simulated in MATLAB [36] using the *GPStuff* [37] which contains an implementation of the quantile regression algorithm used in [32]. Markov-Chain Monte Carlo (MCMC) methods are required to estimate this model. It is well known that MCMC routines carry a large computational burden and the size of the MBTA data set makes this approach infeasible, instead a smaller sample of training data was sampled uniformly across the range of the data. For the purposes of illustration a second experiment was conducted using the MBTA data where a test set of three random days in May was chosen. This was done to allow complete journeys of a bus to be plotted, which is not possible when choosing a random set of observations as a test set due to the extremely low probability of selecting all observations related to one bus journey by chance. Performance will be assessed using the much larger test set of random samples however this 3 day test set allows us to plot a set of predictions from the first prediction to the actual arrival of the bus at the stop and the associated PI to give an intuitive feel for what our approach does in practice. This is not required for the MIT shuttle bus as due to the smaller data set size we divide into a week of training and a single day of test data and so are able to plot complete journeys for the test day.

4.1 MBTA Route 1

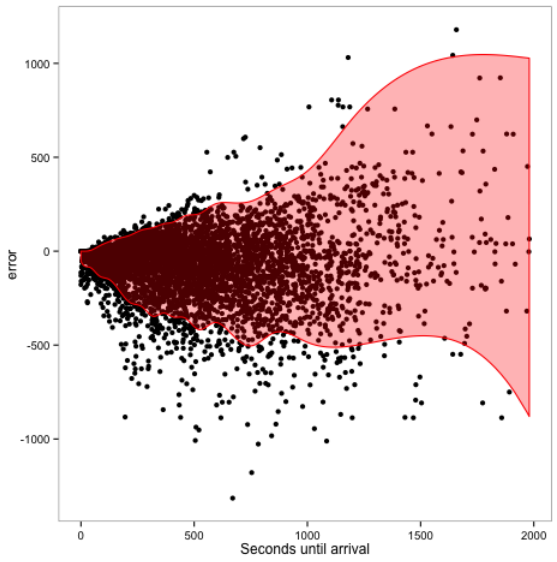
Figure 4 shows the results of the splines models trained using the MBTA data to predict the errors associated with the predicted arrival times test set. The x-axis is the predicted time to arrival in seconds and the y-axis is the error between the predicted and observed arrival times. The 97.5% and 7.5% quantiles estimated using a splines model with 20 degrees of freedom are plotted on top of the data in red. This is contrasted with Figure 4b in which a splines model with 20 degrees of freedom has been used to estimate the median or 50% quantile and then the standard deviation of the residuals used to estimate a confidence interval of 90% as $+/-1.65\sigma$. In both figures and in those shown later the PI is represented as a light red filled in envelope between the upper and lower bounds. While the two approaches both produce intervals that cover 90% of the data the differences are immediately apparent from visual inspection. In the heteroscedastic approach in Figure 4a the greater uncertainty in higher predictions is captured by the inflated envelope after the 1000 second mark, in contrast the bounds up to 500 seconds are much tighter reflecting that predictions made in this region suffer from much less uncertainty. Figure 4b expresses the limitations of the homoscedastic approach, much greater uncertainty is attributed to short range predictions than is true while the bounds on long range predictions exhibit too much confidence.

Figure 5 presents the results for the more computationally intensive GPQR approach. Here a GP is used to learn the upper and lower bounds of the PI. However due to the greater computational burden associated with training a GP a much smaller training data set was sampled uniformly across the range of the data. This is shown in Figure 5a and provides a clear illustration of the heteroscedasticity present in the data. The performance of the GPQR model on a test set of 10,000 samples is shown in Figure 5b. The characteristics observed in the Figure 4a are observed again with the bounds much tighter on predictions less than 500 seconds and inflating more as the uncertainty increases with the size of the prediction. The performance of the GP and splines models are evaluated using PICP and MPIL measures in Tables 2 and we see that even with a much smaller training set the GPQR bounds have a smaller MPIL reflecting on average tighter intervals for the same coverage probability, justifying the additional complexity. The values obtained for the homoscedastic splines model are also provided and they show a much greater MPIL score, however the main result is the observed inability of the homoscedastic approach to adapt to regions of higher and lower uncertainty as illustrated in Figure 4.

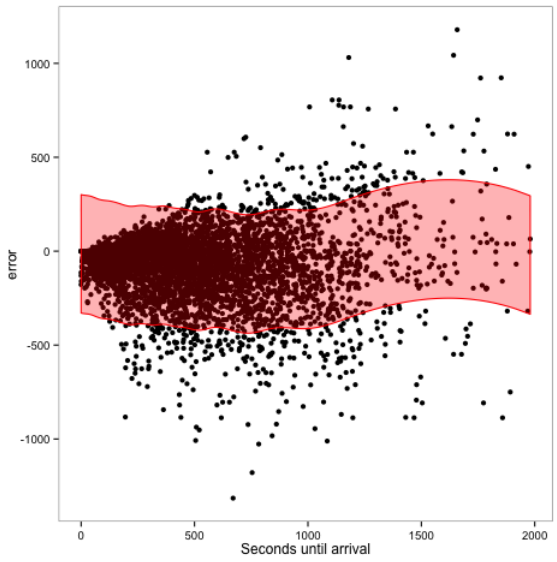
	Test Random		Test Day	
Model	PICP	MPIL	PICP	MPIL
Splines QR	0.89	547.0	0.88	457.14
GP QR	0.88	513.5	0.88	431.1
Splines	0.90	630.3	0.90	700.9

TABLE 2: MBTA Results. The performance of the two different approaches to estimating the quantiles are compared in terms of coverage, PICP and size of interval MPIL on both the random test data and days from May test set. In both cases the GP provides comparable coverage using tighter intervals although the difference is not large.

The performance of the GP and splines approach was also compared in the second experiment conducted on the test set of three complete days in May and is provided in Table 2. Given the superior performance of the GP only the results obtained using this approach will be discussed, these are shown in Figure 6 and are presented in a different

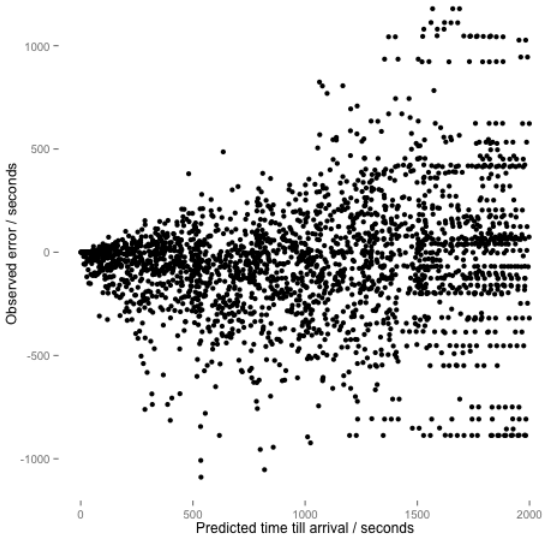


(a) Splines - quantile regression

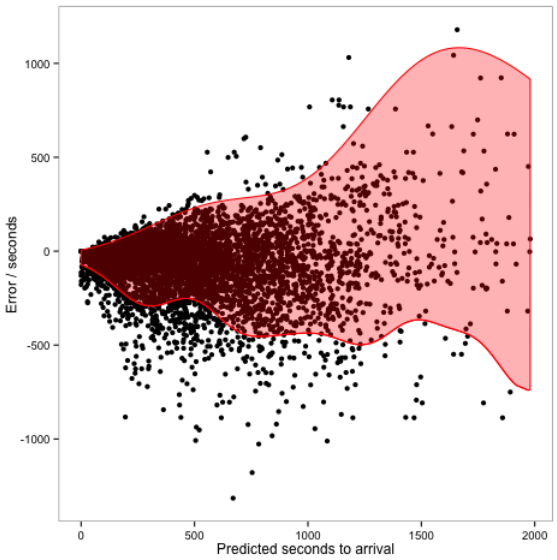


(b) Splines - constant variance

Fig. 4: Comparison of bounds predicted by heteroscedastic, Figure 4a, and homoscedastic, Figure 4b splines models for MBTA test data. The data is plotted with the predicted seconds till arrival on the x-axis and the error between predicted and observed arrival time on the y-axis. The bounds are plotted in red with the envelope defining the PI or CI. Comparison of the intervals highlights the effect of heteroscedasticity with much less uncertainty associated with predictions closer to the arrival time, however the constant variance approach is unable to capture this behaviour.

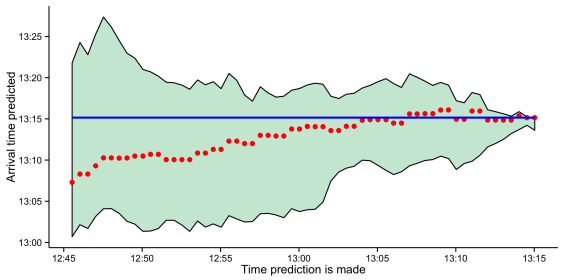


(a) QR Gaussian Process - training set

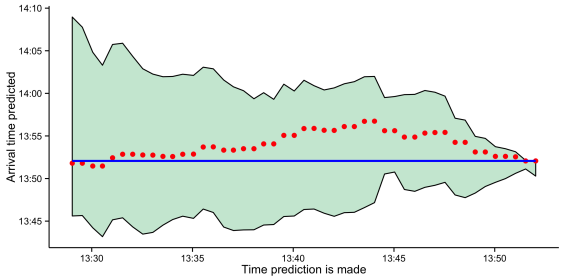


(b) QR Gaussian Process - test set

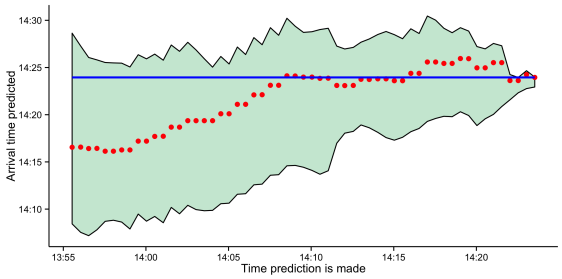
Fig. 5: MBTA data using Gaussian Process model. Figure 5a shows the data used to train the GP quantile regression models. The data was sampled by dividing the range into eighths and taking an equal number of samples from each to minimise bias. Figure 5b shows the PI estimated on the test data which are tighter for smaller predictions but much higher for longer predictions for which there is much more uncertainty.



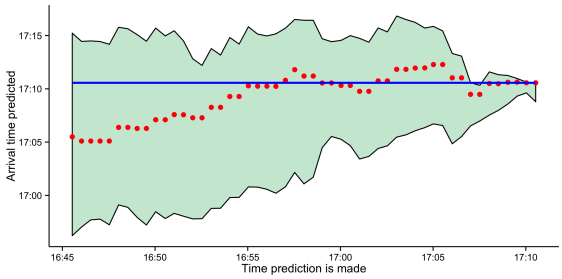
(a) MBTA GPQR - Example 1



(b) MBTA GPQR - Example 2



(c) MBTA GPQR - Example 3



(d) MBTA GPQR - Example 4

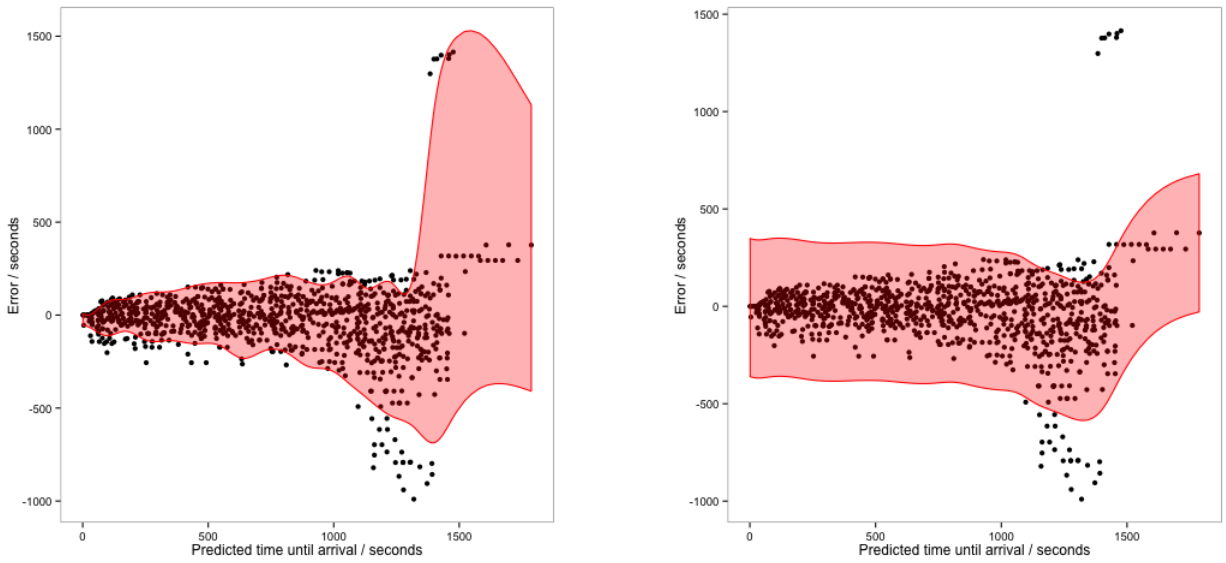
Fig. 6: GP quantile regression results from MBTA full days test data. A number of snapshots are provided showing complete sets of predictions, from first prediction to arrival, to illustrate how the PIs evolve as the bus transits the route. The true fixed arrival time of the bus is shown in blue, with the Nextbus predictions, made every 30 seconds, shown in red. The filled in area represents the interval associated with each prediction, defined by the upper and lower bounds estimated by quantile regression. The x-axis defines the time the prediction is made, while the y-axis defines the actual times predicted.

manner to the first experiment. In order to provide a more intuitive feel for what we are trying to achieve four selected snapshots from this test data are plotted where single complete laps of a bus from earliest prediction to actual arrival are shown. These Figures best illustrate the working of the prediction interval with the time the prediction is made on the x-axis and the predicted time of arrival on the y-axis, the true observed time is plotted as a constant blue line and the predictions are shown as red points that occur every 30 seconds. The PI associated with each prediction is shown as a green envelope who's size varies with the time prediction is made. Closer to the time of arrival we see much more confidence in our predictions. In the longer range predictions the difference between predicted and observed arrival times can be quite significant as much as ten minutes, in other cases prediction times exceed the observed time and result in a situation where a traveller following this advice would miss their bus. It can be observed that in all cases the PI encapsulates the true arrival time regardless of whether the predictions over or underestimate the true arrival time. We now study the much simpler MIT shuttlebus case.

4.2 MIT Shuttlebus

Figure 7 plots the results for the MIT shuttle bus route data set using a week's worth of data for training and an independent test set consisting of a day's worth of observations. Figure 7a shows the test set results obtained using a splines based quantile regression approach which is compared with a homoscedastic splines model in Figure 7b which predicts the median value and estimates a 90% confidence interval as $+/-1.65\sigma$. The constant variance approach of Figure 7b greatly overestimates the uncertainty in the predictions as evidenced by the large distance between the bounds of the confidence interval and the actual observations. In contrast the QR model fits a tight bound in regions of less uncertainty and then inflates these bounds after the 1000 second mark where the uncertainty is much greater. Just as in the previous data sets the performance of the GPQR was compared with that obtained for the splines model however in this case there was no improvement in performance and the faster splines model was preferred, this result is discussed in more detail in Section 5.

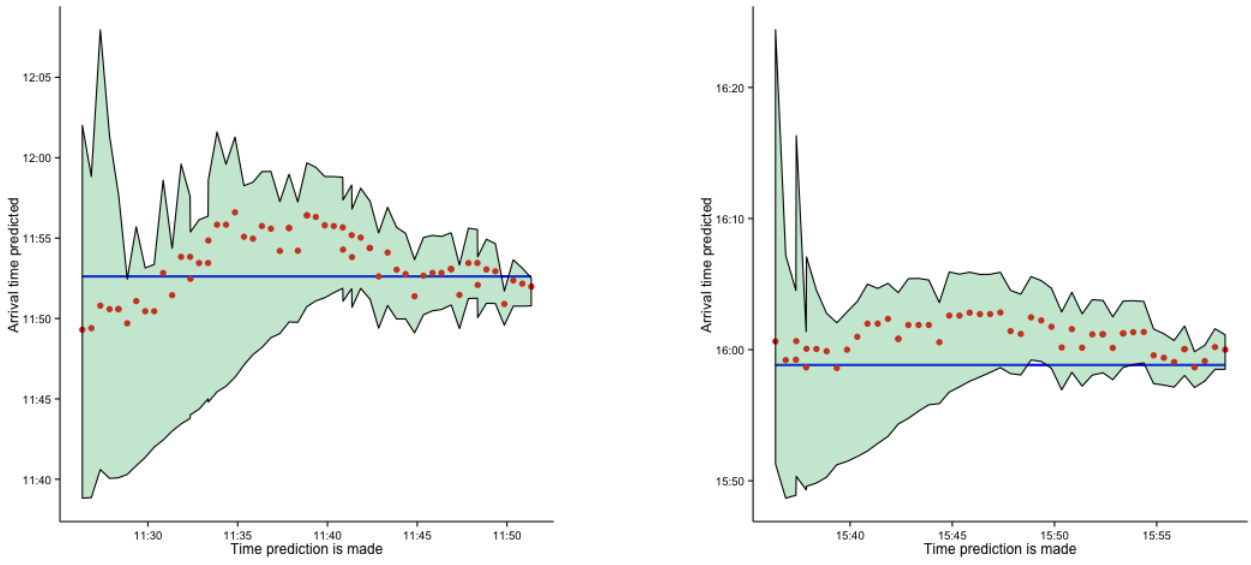
To conclude this section we provide two selected results from the test set that best illustrate the advantages of a prediction interval. These are plotted in Figure 8a and 8b which are now described in detail. In Figure 8a, which is taken from the morning period, the initial bus arrival times predicted are incorrect and 15 minutes earlier than the observed arrival time. However the uncertainty associated with these predictions, which can be inferred from the area of the shaded region, is high which could communicate that these predictions should not be relied upon. Later predictions exhibit much less uncertainty expressing the confidence in the arrival times predicted. This communication of reliability can be of great benefit to commuters allowing them to make informed alternative choices of transport if the situation dictates that it is of high importance to arrive at their destination within a certain timeframe.
Figure 8b shows an example of a shuttlebus lap in the afternoon. It can be seen that the predictions made overestimate the arrival time, meaning that at the predicted time the bus would have already departed. This is a drawback



(a) Quantile regression splines model

(b) Homoscedastic splines model

Fig. 7: Comparison of bounds predicted by heteroscedastic, Figure 7a, and homoscedastic, Figure 7b, splines models for MIT shuttle bus test data. It can be seen that the homoscedastic model greatly overestimates the uncertainty in predictions less than 500 seconds while the bounds predicted for the heteroscedastic model are tight in this region and then inflate as the uncertainty increases as the predicted time till arrival becomes larger.



(a) Morning

(b) Afternoon

Fig. 8: Selected results from the test data for MIT Boston daytime shuttlebus. The true fixed arrival time of the bus is shown in blue, with the Nextbus predictions, made every 30 seconds, shown in red. The filled in area represents the interval associated with each prediction, defined by the upper and lower bounds estimated by quantile regression. The x-axis defines the time the prediction is made, while the y-axis defines the actual times predicted. Figure 8a shows a single lap of the shuttlebus in the morning. Figure 8b shows a lap of the shuttlebus this time in the afternoon. See text for detailed discussion.

of using a single valued prediction. By contrast we see that the true arrival time falls within the prediction interval. The implications of these results and some possible areas for further exploration will be discussed in the next section.

5 Discussion

In this paper we have described a 'meta-model' approach that can be applied to existing predictive systems to handle the uncertainty associated with bus arrival time predictions. This process treats the algorithm generating the predictions as a black-box which has the advantage of being agnostic to the process generating the predictions, meaning it can be deployed on any system without the need for alteration of the existing algorithm. However it must be emphasised that our approach is not intended for or capable of improving the quality of the predictions made, merely estimating the uncertainty associated. The approach was demonstrated using real world data from a TIS in Boston. Two examples were utilised; one of a busy main route through the city and the second showing the simplest possible setting of a single bus on loop. In both cases statistical analysis has revealed that the uncertainty in the predictions exhibits heteroscedasticity. This is an important finding that has not previously been made in the literature, although anecdotal evidence, as in [15], certainly points to the presence of heteroscedasticity in other data. As we have stated the meta-model approach is applicable to any existing predictive system. As well as this the tools used for analysis are all freely available online and we have provided links to both the R package required for quantile regression *quantreg* [35] and the MATLAB toolbox *GPstuff* [37] that allows GPQR to be deployed. The approach used here is therefore both portable and replicable and we hope that other researchers will take these tools and build on the meta-model methodology outlined here applying them to their own data.

With the presence of heteroscedasticity in the predictions established an approach to appropriately handle this effect was developed using quantile regression to learn upper and lower bounds on the observed error. The PIs learned in this way were demonstrated to provide the desired coverage on unseen test data and the resulting visualisations illustrate the characteristic heteroscedastic behaviour of changing levels of uncertainty associated with predictions made at different times. The capability to provide this envelope within which the bus can be guaranteed to arrive with a given probability is another contribution made here and the meta-model approach used means that it can be applied to existing TIS systems. Another advantage of the PI approach is that it more easily allows for asymmetry in the bounds in contrast to the homoscedastic method, this was demonstrated by our choice of quantiles, the upper bound was set to 97.5% and the lower to 7.5% to reflect the fact that the errors observed were generally biased towards under estimating the arrival time. This can be adapted either way and it is of course possible to obtain much smaller intervals using lower PIs or to penalise the possibility of missing the bus by skewing the PI in the opposite manner. The results from using a PI composed of the 95% and 5% quantiles produced a larger interval for the same coverage.

A comparison of two approaches for the learning of these PIs was performed on the data to ascertain what the superior method was: the standard splines QR or the more advanced and recently developed GPQR . Interpretation of these results is aided by some theoretical background on the methods. A Gaussian Process is a stochastic process that is fully specified by a mean function and covariance function and is a highly flexible non-parametric approach. While cubic splines regression was developed independently in a different context works such as [23] have shown that the method is equivalent to a GP with a more restricted covariance function. Thus the GP can be considered as a more flexible general model. This is borne out in the results obtained where the GPQR approach matches or exceeds results obtained using splines QR. Thus it would seem that one should always make use of GPQR. However, one important point for researchers to consider is that this additional flexibility carries a heavy computational cost with the GPQR requiring much greater computational resources to estimate in comparison to the splines model. Therefore judgement is required as to whether the situation is complex enough to call for the more flexible approach. We found there to be no improvement over the splines model in modelling PIs for the MIT shuttlebus data, which consits of a single bus on loop over a much shorter route than the MBTA data.

We endeavoured to demonstrate the practical usefulness of these methods through a number of illustrative figures that zoomed in on the set of predictions throughout the complete journey of a bus from first prediction to arrival in Figures 6 and Figure 8. The variability in the predictions and the errors illustrated are a clear problem for travellers aiming to plan their journeys reliably and the use of PIs in this setting can restore confidence in shorter range predictions whilst also allowing travellers to take into account the degree of uncertainty associated with longer range predictions and adapt their plans accordingly.

In the future we aim to develop this approach to incorporate features that give more information as to the context at the time of prediction such as weather etc. This should allow us to obtain more optimal PIs that are narrower for conditions of less uncertainty and therefore of greater practical use to the public.

References

[1] T. R. C. Programs, "Real time bus arrival information systems," *TRCP Synthesis*, vol. 48, pp. 1–71, 2003.

[2] L. Tang and T. Piyushimita, "An analysis of anticipated behavioral responses to real-time transit information systems," *World Conference on Transport Research Society*, vol. 11, 2007.

[3] L. Tang and P. Thakuriah, "Ridership effects of real-time bus information system: A case study in the city of chicago," *Transportation Research Part C*, vol. 22, pp. 146–161, 2012.

[4] F. Zhang, Q. Sheng, and K. J. Clifton, "Examination of traveler responses to real-time information about bus arrivals using panel data," *Transportation Research Record: Journal of The Transportation Research Board*, vol. 2082, pp. 107–115, 2008.

[5] E. Mazloumi, G. Rose, G. Currie, and M. Savri, "An integrated framework to predict bus travel time and its variability using traffic flow data," *Journal of Intelligent Transportation Systems*, vol. 15, no. 2, pp. 75–90, 2011.

[6] M. Altinkaya and M. Zontul, "Urban bus arrival time prediction: A review of computational models," *International Journal of Recent Technology and Engineering*, vol. 2, pp. 164–169, 2013.

[7] S. Zadeh, T. Awar, and M. Basirat, "A survey on application of artificial intelligence for bus arrival time prediction," *Journal of Theoretical and Applied Information Theory*, vol. 46, pp. 516–525, 2012.
[8] M. Zaki, I. Ashour, M. Zorkany, and B. Hesham, "Online bus arrival time prediction using hybrid neural network and kalman filter techniques," *International Journal of Modern Engineering Research*, vol. 3, pp. 2035–2041, 2013.
[9] M. Yang, Y. Liu, and Z. You, "The reliability of travel time forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, pp. 162–171, 2010.
[10] J. Bates, J. Polak, P. Jones, and A. Cook, "The valuation of reliability in personal travel," *Transport Research Part E: Logistics and Transportation Review*, vol. 37, pp. 191–229, 2001.
[11] C. Sun, G. Arr, and R. P. Ramachandran, "An investigation in the use of vehicle reidentification for deriving travel time and travel time distributions," *Transportation Research Record: Journal of The Transportation Research Board*, vol. 1826, pp. 25–30, 2003.
[12] Z. Li, D. A. Hensher, and J. M. Rose, "Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence," *Transport Research Part E: Logistics and Transportation Review*, 2010.
[13] E. Mazloumi, Currie, and Rose, "Using gps data to gain insight into public transport travel time variability," *Journal of Transport Engineering*, vol. 136, pp. 623–631, 2010.
[14] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. W. C. van Lint, "Prediction intervals to account for uncertainties in travel time prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 537–547, 2011.
[15] C. Coffey, A. Pozdnoukhov, and F. Calabrese, "Time of arrival predictability horizons for public bus routes," *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Computational Transportation Science*, pp. 1–5, 2011.
[16] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of Economic Perspectives*, vol. 15, pp. 143–156, 2001.
[17] F. C. Pereira, C. Antoniou, J. Aguilar, and M. Ben-Akiva, "A metamodel for estimating error bounds in real-time traffic prediction systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1310–1322, 2014.
[18] www.nextbus.com, 2014.
[19] T. Breusch and A. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979.
[20] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980.
[21] C. W. J, J. M. E, and J. M. M, "A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data." *Technometrics*, vol. 23, pp. 351–361, 1981.
[22] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1107, 1982.
[23] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. MIT press, 2006.
[24] S. Weisberg, *Applied Linear Regression*. Wiley, 2005.
[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[26] R. Koenker, *Quantile Regression*. Cambridge, 2005.
[27] T. C. Chiang and J. Li, "Stock returns and risk: Evidence from quantile regression analysis," *Journal of Risk and Financial Management*, vol. 5, pp. 20–58, 2012.
[28] M. Geraci and M. Bottai, "Quantile regression for longitudinal data using the asymmetric laplace distribution," *Biostatistics*, vol. 8, pp. 140–154, 2006.
[29] J. Yang, X. Meng, and M. W. Mahoney, "Quantile regression for large-scale applications," *International Conference of Machine Learning*, vol. 28, pp. 1–8, 2013.
[30] K. Yu and R. A. Moyeed, "Bayesian quantile regression," *Statistics and Probability Letters*, vol. 54, pp. 437–447, 2001.
[31] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola, "Nonparametric quantile estimation," *Journal of Machine Learning Research*, vol. 7, pp. 1231–1264, 2006.
[32] A. Boukouvalas, R. Barillec, and D. Cornford, "Gaussian process quantile regression using expectation propogation," *International Conference of Machine Learning*, pp. 1–8, 2012.
[33] T. P. Minka, "Expectation propagation for approximate bayesian inference," *Conference on Uncertainty in Artificial Intelligence*, pp. 362–369, 2001.
[34] R. C. Team, "R: A language and environment for statistical computing," *R Foundation for Statistical Computing, Vienna, Austria*, 2013.
[35] R. Koenker, "Quantile regression in r: A vignette," *Online*, 2012.
[36] Mathworks. (2013) Matlab and statistics toolbox release 2013a. Inc. Natick, Massachusetts, U.S.
[37] J. Vanhatalo, J. Riihimaki, J. Hartikainen, P. Jylanki, V. Tolvanen, and A. Vehtari, "Gpstuff: Bayesian modeling with gaussian processes," *Journal of Machine Learning Research*, vol. 14, pp. 1175–1179, 2013.

Aidan O'Sullivan received his PhD. in Statistics from the Department of Mathematics in Imperial College London in 2013. He is currently a postdoctoral associate in the Future Urban Mobility group in the Singapore MIT Alliance for Research and Technology. He holds a Bachelor's degree in Electrical and Electronic Engineering from University College Cork and a MSc. in Bio-Engineering from Imperial College London.

Francisco C. Pereira Francisco C. Pereira is Full Professor at the Technical University of Denmark (DTU), where he leads the ITS research group. Previously, he was Senior Research Scientist at MIT/CEE ITSLab, where he worked both in Boston and Singapore, as part of the Singapore-MIT Alliance for Research and Technology, Future Urban Mobility project (SMART/FM).

Jinhua Zhao Jinhua Zhao is the Edward H. and Joyce Linde Assistant Professor in the Department of Urban Studies and Planning at MIT. He holds Master of Science, Master of City Planning and Ph.D. degrees from MIT and a Bachelor's degree from Tongji University. He studies 1) behavioral foundation for transport policies; 2)public transit management; and 3) China's urbanization and urbanmobility. Prof. Zhao directs the urban mobility lab JTL and is the co-PI for the Transit Lab at MIT.

Haris Koutsopoulos Haris N. Koutsopoulos is Professor in the Department of Civil and Environmental Engineering at Northeastern University in Boston and Guest Professor at the Royal Institute of Technology in Stockholm. His research has focused on the modelling of Intelligent Transportation Systems, traffic simulation models at various levels of resolution, and methods and algorithms for their calibration. His current research interests are in the use of data from opportunistic and dedicated sensors to improve planning, operations, and monitoring and control of urban transportation systems. He founded the iMobility lab focused on the use of Information and Communication Technologies to address urban mobility problems. The lab received the IBM Smarter Planet Award in 2012.